TriSum: Learning Summarization Ability from Large Language Models with Structured Rationale

Pengcheng Jiang[1], Cao Xiao[2], Zifeng Wang[1], Parminder Bhatia[2], Jimeng Sun[1], Jiawei Han[1]

[1] University of Illinois Urbana-Champaign [2] GE HealthCare

NAACL'24

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

GE HealthCare

Introduction

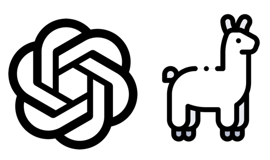
In the era of information overload, text summarization has become a crucial tool for quickly grasping the essence of lengthy documents.

Our focus: Abstractive Summarization

Small PLMs (BERT/BART/T5)

Pros: Cost-effective for fine-tuning;

Cons: Low factualness and interpretability

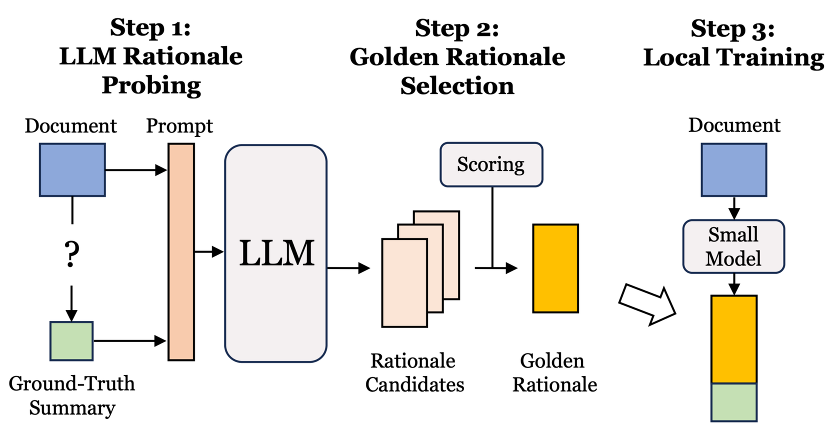


LLMs (e.g., GPT, LLaMA)

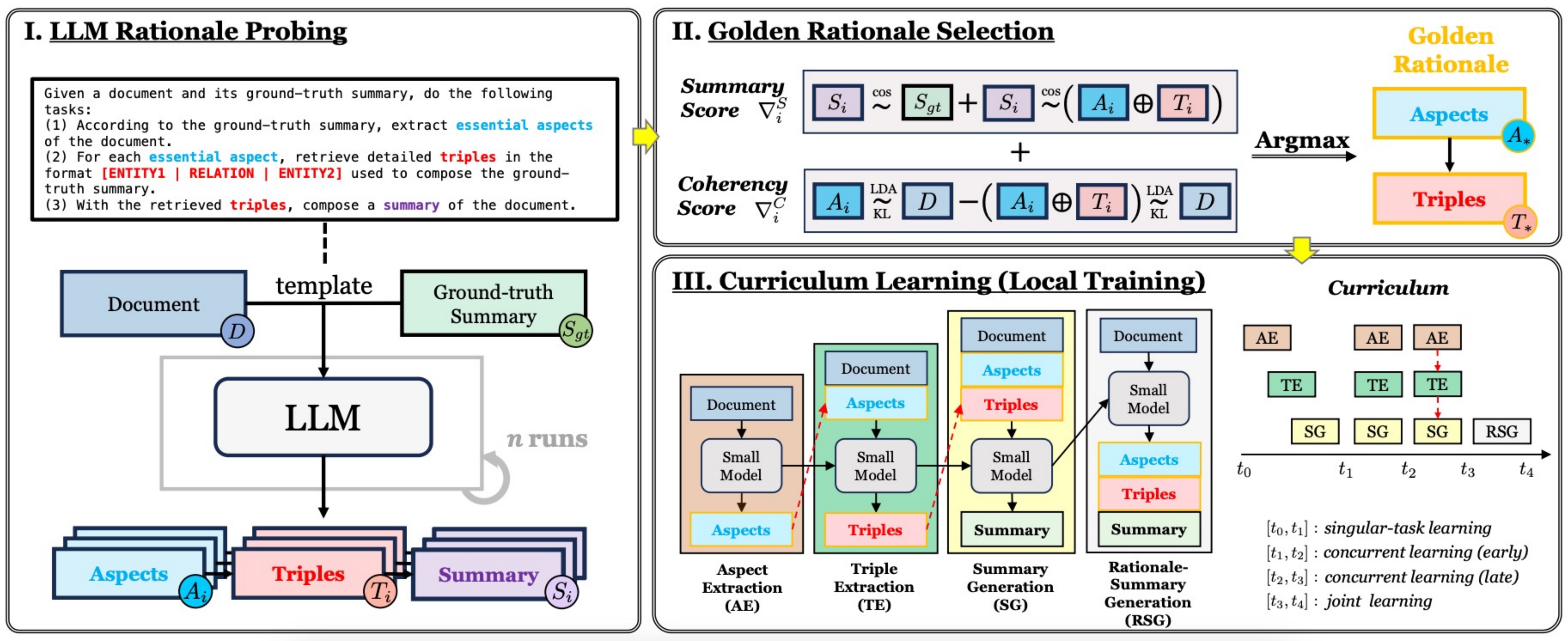
Pros: High interpretability with rationale; High NLU capabilities.

Cons: Costly for fine-tuning

Can we train a small model to learn the interpretable summarization ability from LLMs?



Methodology – TriSum



TriSum Framework

Step 1 – LLM Rationale Probing:
For each pair of <document, ground-truth summary>, we let the LLM generate essential aspects, relationship triples, and a summary, as a structured rationale.

Step 2 – Golden Rationale Selection:
Summary Score: evaluates the semantic similarity between the generated summary and the ground truth. **Coherence Score**: measures how well the aspects and triples align with the document's latent topics.

Step 3 – Local Training:
We employ a ***curriculum learning*** strategy, starting with simpler tasks to the more complex task of a rationale-summary generation.

$$\mathcal{L}_A = -\sum \log p(A_*|D;\theta_s),$$
$$\mathcal{L}_T = -\sum_{D\in\mathcal{D}} \log p(T_*|D,A_*;\theta_s),$$
$$\mathcal{L}_S = -\sum_{D\in\mathcal{D}} \log p(S_{gt}|D,A_*,T_*;\theta_s).$$

$$\mathcal{L}_{\text{concurrent-late}} = -\sum_{D\in\mathcal{D}}\Big[\log p(A_*|D;\theta_c) + \log p(T_*|D,\tilde{A};\theta_c) + \log p(S_{gt}|D,\tilde{A},\tilde{T};\theta_c)\Big],$$

$$\mathcal{L}_{\text{joint}} = -\sum_{D\in\mathcal{D}}\Big[\lambda_R \log p(R_*|D;\theta_r) + \lambda_S \log p(S_{gt}|D,\tilde{R};\theta_r)\Big],$$

Experiments & Results

	CNN/DailyMail				XSum				ClinicalTrial			
Model	**R-1**	**R-2**	**R-L**	**Δ**	**R-1**	**R-2**	**R-L**	**Δ**	**R-1**	**R-2**	**R-L**	**Δ**
Baselines												
BERTSumAbs (Liu and Lapata, 2019)	41.2	18.7	37.2	+13.6%	38.8	16.5	31.0	+28.3%	39.2	19.3	29.6	+19.3%
$T5_{Large}$ (Raffel et al., 2020)	42.4	20.8	39.9	+7.0%	40.1	17.2	32.3	+23.5%	41.3	22.1	32.5	+9.6%
$BART_{Large}$ (Lewis et al., 2019)	44.0	21.1	40.6	+4.4%	45.4	22.3	37.3	+5.4%	43.5	23.3	33.7	+4.6%
PEGASUS (Zhang et al., 2020)	44.2	21.6	41.3	+3.0%	**46.7**	**24.4**	38.9	+0.6%	41.8	22.9	31.7	+9.0%
GSum (Dou et al., 2021)	45.5	22.3	**42.1**	+0.4%	45.1	21.5	36.6	+7.3%	43.5	23.1	32.8	+5.7%
$BigBird_{Large}$ (Zaheer et al., 2021)	43.8	21.1	40.7	+4.5%	47.1	24.1	38.8	+0.6%	**44.2**	23.8	**34.5**	+2.5%
SimCLS (Liu and Liu, 2021)	45.6	21.9	41.0	+1.7%	46.6	**24.2**	**39.1**	+0.7%	43.8	23.3	34.1	+3.9%
SeqCo (Xu et al., 2022)	45.0	21.8	41.8	+1.6%	45.6	22.4	37.0	+5.4%	42.8	22.5	33.2	+6.7%
$GLM_{RoBERTa}$ (Du et al., 2022)	43.8	21.0	40.5	+4.7%	45.5	23.5	37.3	+4.1%	43.3	23.0	33.9	+4.9%
$GPT\text{-}3.5_{zero\text{-}shot}$	37.4	13.8	29.1	+37.4%	26.6	6.7	18.8	+112.5%	34.8	12.8	23.5	+47.8%
Our Method												
GPT-3.5 w/ TriSum rationale	**46.7**	**23.5**	40.7	−0.5%	34.4	12.6	28.4	+46.8%	**44.6**	**24.5**	30.4	+5.6%
TriSum-S	**45.9**	**22.8**	**42.3**	−0.6%	**47.4**	**24.8**	**39.4**	−1.0%	**45.3**	**24.8**	**35.0**	+0.0%
TriSum-C	45.5	22.3	41.2	+1.2%	46.5	24.0	38.7	+1.1%	**44.2**	23.7	34.4	+2.7%
TriSum-J	**45.7**	**22.7**	**41.9**	—	**47.3**	**24.4**	**39.0**	—	**45.3**	**24.6**	**35.2**	—

ROUGE score performance

	CNN/DailyMail		XSum		ClinicalTrial	
Model	**BS**	**BAS**	**BS**	**BAS**	**BS**	**BAS**
Baselines						
BERTSumAbs	85.76	-3.81	87.23	-3.66	85.41	-3.79
$T5_{Large}$	87.22	-3.71	90.73	-2.70	87.76	-2.89
$BART_{Large}$	87.98	-3.45	91.62	-2.50	88.30	-2.79
PEGASUS	87.37	-3.64	91.90	-2.44	87.62	-2.80
GSum	87.83	-3.54	91.23	-2.57	88.41	-2.75
$BigBird_{Large}$	88.03	-3.38	**91.97**	**-2.40**	**89.45**	-2.67
SimCLS	88.28	-3.39	90.78	-2.93	87.85	-3.15
SeqCo	87.47	-3.56	91.35	-2.56	88.06	-2.93
$GLM_{RoBERTa}$	87.33	-3.69	91.87	-2.51	88.55	-2.84
$GPT\text{-}3.5_{zero\text{-}shot}$	87.70	-3.36	87.67	-2.80	87.08	-3.01
Our Method						
$GPT\text{-}3.5^{*}_{TriSum}$	**89.20**	**-3.14**	89.25	-2.58	89.20	**-2.55**
TriSum-S	**88.48**	**-3.22**	**91.95**	**-2.38**	**90.05**	**-2.47**
TriSum-C	87.21	-3.76	90.88	-2.84	89.40	-2.59
TriSum-J	**88.50**	**-3.25**	**92.17**	**-2.33**	**89.97**	**-2.53**

BERTScore/BARTScore performance

Findings:

1. On average, TriSum-J consistently outperformed state-of-the-art baselines, achieving improvements of 4.5%, 8.5%, and 7.4% in ROUGE scores, respectively.
2. TriSum-generated rationales also enhanced the performance of LLMs themselves (GPT-3.5 w/ TriSum's rationale), showcasing the value of interpretable rationales.

Article:
(CNN) Four workers died in a massive oil rig fire that raged for hours off the coast of Mexico Wednesday. Mexican state oil company Pemex said 45 workers were injured in the blaze, which began early Wednesday morning. Two of them are in serious condition, the company said. Authorities evacuated about 300 people from the Abkatun Permanente platform after the fire started, Pemex said. At least 10 boats worked to battle the blaze for hours. The fire had been extinguished by Wednesday night, Pemex said in a Twitter post. The company denied rumors that the platform had collapsed and said there was no oil spill as a result of the fire. The state oil company hasn't said what caused the fire on the platform, which is located in the Gulf of Mexico's Campeche Sound. The fire began in the platform's dehydration and pumping area, Pemex said. CNN's Mayra Cuevas contributed to this report.

Ground truth summary:
The fire on a platform in the Gulf of Mexico has been extinguished, Pemex says . 45 workers were injured in the blaze, according to the state oil company . Four workers were killed in the oil rig fire, which started early Wednesday .

BART summary:
About 300 people were evacuated from the Abkatun Permanente platform. The fire began in the platform's dehydration and pumping area, Pemex says. The company denies rumors that the platform had collapsed and said there was no oil spill as a result of the fire.

TriSum rationale:
<rationale> <aspects> Massive oil rig fire off the coast of Mexico. Four workers died. 45 workers injured. Two workers in serious condition. 300 people evacuated. Fire extinguished by Wednesday night. No oil spill. **<triples>** [Four workers | were died in | oil rig fire], [45 workers | were injured in | the blaze], [Two workers | are in | serious condition], [300 people | were evacuated from | the platform], [The fire | had been extinguished by | Wednesday night], [The fire | did not result in | oil spill].

TriSum summary:
Four workers were died in an oil rig fire. 45 workers were injured, two of them in a serious condition, state oil company Pemex says. About 300 people were evacuated after the fire started early Wednesday. The fire has been extinguished and it did not result in oil spill, the company says.

An example of abstractive summarization of an article in CNN/DailyMail dataset.
We use different colors to show the distinct topics in the article and summary.

The rationale provides a structured breakdown of the essential information, enhancing the interpretability of the summarization process.

Future Works

1. Adapting TriSum to other NLP tasks (e.g., QA, machine translation)
2. Generating richer rationales with graph representations or knowledge bases
3. Developing interactive, user-centric summarization systemsuser-centric experiences.

…

Thank you for your interest!
Please email pj20@illinois.edu for any questions.